DOCUMENT RESUME

ED 455 253 TM 033 055

AUTHOR Lee, Yong-Won
TITLE The Essay Scoring and Scorer Reliability in TOEFL CBT.
PUB DATE 2001-04-00
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Computer Assisted Testing; *English (Second Language); *Essay Tests; *Interrater Reliability; Language Tests; Online Systems; Psychometrics; Scores; *Scoring; *Writing Tests
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

An essay test is now an integral part of the computer based Test of English as a Foreign Language (TOEFL-CBT). This paper provides a brief overview of the current TOEFL-CBT essay test, describes the operational procedures for essay scoring, including the Online Scoring Network (OSN) of the Educational Testing Service (ETS), and discusses major psychometric issues related to the reliability of essay scores. Test takers can type or hand-write their written responses to the given essay prompt, and hand-written essays are scanned. All the writing samples are electronically distributed to trained ETS readers at regional sites through the OSN. The essays are scored online, and the scores are posted to the central database for use in score reporting. The two major sources of potential scoring error may come from the raters and the prompts. Readers undergo strict training, and prompts are reviewed thoroughly by content experts and monitored in terms of their performance through the OSN. It would be difficult to compute inter-reader reliability indices by using individual readers as units of reliability analysis, but rater reliability can be computed based on ratings, not readers, so that inter-reader reliability can possibly be examined at the levels of relative ratings, absolute ratings, and absolute readers. More research is necessary on the comparability of writing prompts. Future avenues for TOEFL-CBT essay scoring are automatic essay scoring and the development of new integrative writing tasks. (Contains 6 figures and 14 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 455 253

The Essay Scoring and Scorer Reliability in TOEFL CBT

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

Y.-W. Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee

Educational Testing Service

Rosedale Road Princeton

NJ 08541

Phone: 609-732-1348 (Office)

Fax: (609)-734-1755

Email: ylee@ets.org

TM033055

Paper presented as a part of the "Computer-Based TOEFL" symposium at the annual meeting of National Council on Measurement in Education (NCME) held in Seattle, Washington between April 11-13, 2001.

BEST COPY AVAILABLE

The Essay Scoring and Scorer Reliability in TOEFL CBT

Objectives

The essay test (formerly known as the Test of Written English) is now an integral part of the computer-based TOEFL test (TOEFL CBT henceforth) administered through the Prometric and ETS CBT test centers world-wide. The TOEFL CBT was first introduced in July 1998 and then rolled out to Asian countries in October 2000. A pool of essay prompts is prepared and rotated off periodically for test security reasons. Now that we have a substantial body of data, we can report on the psychometric quality and actual performance of this CBT component to the language testing community. The purpose of this presentation is to (1) provide a brief overview of current TOEFL CBT essay test, (2) describe operational procedures for essay scoring including the Online Scoring Network (OSN) at Educational Testing Service (ETS), and (3) discuss major psychometric issues related to the reliability of essay scores.

Brief History and Background: CBT Writing Tests

Prior to the introduction of TOEFL CBT, the Test of Written English (TWE) had been offered as a required component of the paper-based TOEFL test in only 5 of 12 administrations a year since1986. The test was designed to complement the Structure and Written Expression (SWE) Section of the paper-TOEFL as a direct measure of writing, but is not included in the SWE section score (Educational Testing Service, 1996). Instead, the TWE score is reported separately on the examinee score report. The TWE continues to be administered in conjunction with the paper-based TOEFL, but only in a limited area, such as the People's Republic of China, because the paper-based TOEFL was replaced by the TOEFL CBT in most parts of the world, as of October 2000.

The TOEFL CBT essay test is being administered as a mandatory component for all TOEFL CBT test takers. In TOEFL CBT, the essay test is given at the end of the test after the 3 multiple-choice (MC) sections (*i.e.*, Listening, Structure, and Reading), and the essay score is now completely integrated into the SWE section score. The essay score accounts for about half of the Section 2 reported score and one sixth of the total score. A writing prompt is pseudo-randomly chosen for each examinee from a pool of prompts that is periodically redefined. Each examinee is given 30 minutes to write his/her response to the given prompt. A large pool of actual writing prompts (about 150) for TOEFL CBT is published in the TOEFL Bulletin so that examinees can have equal chances of familiarizing themselves with the prompts in advance.

Examinees can choose to either type or hand-write their written responses to the given prompt. Hand-written essays are scanned. All the writing samples are electronically distributed to trained ETS readers through the OSN network (In this paper, readers and raters will be used interchangeably with the same meaning). CBT essays are rated holistically according to a 6-point rating scale. More detailed information about the

scoring rubric can be found in the Computer-Based TOEFL Score User's Guide (ETS, 1998) and the TOEFL CBT Bulletin (ETS, 2000). Off-topic essays are not scored, but are given a score of 0. The sum of the estimated number-right score on the Structure section and the weighted essay rating is converted to a scale score ranging from 0 to 30.

Essay Scoring: The Online Scoring Network (OSN).

The OSN system captures essay responses from the CBT system, stores them in a central database, and serves them to trained ETS readers located in regional sites. Then the essays are scored online, and the scores then posted to the central database for use in score reporting. ETS trained readers can access examinee essays through OSN. Among the essays rated by ETS readers, there are always pre-scored *monitor essays* included so that individual reader performance may be monitored on an on-going basis.

To be qualified as readers for OSN, all applicants must receive training and pass a certification test to show that they have mastered the OSN software and the scoring procedures used for the particular test(s) of interest (Powers & Kubota, 1998). All the essays are rated independently according to a 6-point scoring rubric by two readers, who were selected from a pool of qualified ETS readers. Reported essays scores are a rounded average of the two ratings in increments of .5 (0.0, 1.0, 1.5, …, 5.5, 6.0). If there is a discrepancy greater than one point, the essay will be rated again by a third rater. The two closest ratings are used to compute the reported score for such cases.

Performance summaries of raters and prompts can be generated through OSN, as shown in Figure 1. The performance summary provides the following information: the total number of essays read, mean scores, score distribution, reading rates, percent of perfect agreement, percent of adjacent ratings, and discrepancy rates during a certain time period. Such kinds of information can be produced for both raters and prompts.

Insert Figure 1 about here

Psychometric Issues Related to Reliability of Essay Scores

In TOEFL CBT essay scoring, the two major sources of potential scoring error may come from raters and prompts. For this reason, readers undergo strict rater training and continuous rater performance monitoring through OSN, and writing prompts are also reviewed thoroughly in many steps by content experts and monitored in terms of their performance through OSN. A third rating is used when unusual discrepancies occur even after all these quality control measures were taken for raters. The reliability of the CBT essay scores can be largely looked at in terms of raters, prompts, and the combination of both.

1. Inter-Reader Reliability.

Because neither a purely crossed nor a nested design is used for examinees and readers in scoring CBT essays, it would be difficult to compute inter-reader reliability indices by using individual readers as units of reliability analysis. Rather, rater reliability can be computed based on ratings (*i.e.*, the first and second ratings), not readers. In this sense, the inter-reader reliability can possibly be examined at three different levels in the context of TOEFL CBT essays: (1) ratings (relative), (2) ratings (absolute), and (3) readers (absolute).

Insert Table 1 about here

First, it is possible to use the ratings as the basic units of analysis for calculating reader reliability for a group of examinees. A correlation between the first and second ratings for a group of essays can be used as a measure of the consistency between the two ratings over a group of examinees from a norm-referenced perspective. The alpha can be also obtained based on this correlation and reported, as in Table 1-b. These indices may be computed by the month and displayed graphically to show a trend in inter-reader reliability over time. As shown in Table 1-b and Figure 2, the inter-reader reliability was 0.84 for the whole year of 1999, but varied from 0.81 to 0.86 from month to month. The correlation between the two ratings was .73 for the whole year and it ranged from 0.69 to 0.76 over 12 months.

Insert Figure 2 about here

Second, a number of agreement indices (*e.g.*, percent of perfect or adjacent score agreement, discrepancy rates, and kappa) can be computed by using the ratings as the unit of analysis as well. A cross-tabulation of the first and second ratings can provide a clear picture of overall reader agreement during a certain time period (See Table 1-a). The correlation and the alpha do not take into account the systematic rating errors between the ratings, but these agreement indices can provide reader reliability information from a criterion-referenced perspective. These indices can be also computed by the month, as in Figure 3.

As shown in Table 1-b, the percents of perfect and adjacent agreement between the two ratings were 53.5% and 42.8%, respectively, for 1999, whereas the discrepancy (more than one point discrepancy) rate was as low as 3.7%. The perfect agreement rate

ranged from 51.2% to 56.4% over 12 months in 1999, while the discrepancy rate ranged from 2.6% to 4.8% (See Figure 3). When the perfect and adjacent agreement rates are combined together, the combined rate exceeded 95% in all 12 months of 1999.

Insert Figure 3 about here

Third, these agreement indices can be computed at the level of each individual reader, as done in the OSN system. The pre-scored *monitor essays* are used to monitor the performance of individual raters against a certain criterion on an on-going basis. Another good example of individual rater indices may be obtained from the many-faceted Rasch analysis: each reader can be calibrated in terms of severity/leniency; within-rater inconsistencies are investigated through rater fit statistics for each individual rater, as demonstrated by Myford, Marr, and Linacre (1996) by using TWE data.

2. Comparability of writing prompts.

The comparability of writing prompts in terms of difficulty can be not only a fairness issue for different examinee subpopulations (Golub-Smith, Reese & Steinhaus, 1993) but also a score reliability issue that can impact the dependability of essay scores in the context of TOEFL. Since only one essay is administered per examinee in TOEFL CBT, however, there is no established way of computing internal consistency among prompts based on the operational essay data or of making statistical adjustments for prompt difficulty difference through equating. It may be only said that essays are linked to one another thorough a non-statistical *social moderation* procedure (Linn, 1993; Mislevy, 1992), because pre-scored *monitor* papers (often called *benchmark* or *anchor* papers) are used to exemplify the rating criteria in rater training and are interspersed with yet-to-be-rated papers as a means of monitoring rater performance.

For this reason, comparability of writing prompts may be examined roughly by looking at the raw essay means of the writing prompts, as shown in Figure 4. However, one may hardly expect that these raw mean scores of different prompts are comparable in a strictest sense of equated scores, because different groups of examinees take a different prompt. To overcome this problem, Breland, Muraki, and Lee (2000) used the combined score of multiple-choice sections of TOEFL CBT as an anchor and used a logistic regression approach to come up with a prompt difficulty index and prompt DIF measure. However, it should be noted that the MC section may not be an ideal anchor for the TOEFL examinees' writing ability, as pointed out by DeMauro (1992). More research might be necessary on this line in the future. As shown in Figure 4, the raw mean scores for prompts ranged from 3.84 to 4.17, with a mean of 4.03 and a standard deviation of 0.06.

Insert Figure 4 about here

Another possible way of monitoring prompt performance might be to look at inter-reader reliability prompt by prompt. Figures 5 and 6 display the distributions of the inter-reader reliability and the agreement indices among more than 80 writing prompts administered in 1999. As shown in Figure 5, the correlation between the two ratings approximately ranged from 0.69 to 0.79 for the prompts, whereas the alpha ranged from 0.81 to 0.88. On the other hand, the agreement rate varied from 49% to 61%, while the discrepancy rate ranged from 2% to 6%.

Insert Figures 5 and 6 about here

---

3. Score reliability of TOEFL CBT essays

A special generalizability-theory study might be necessary to estimate the relative effects of essay prompts and raters on the essay score. As mentioned earlier, it is not possible to obtain any internal consistency reliability coefficient based on the operational CBT essay data. In 1999, more than 80 writing prompts were assigned to examinees in a near-random way, and each essay was given two readers' ratings. It may be possible to simply treat it as if examinees (p) were nested within prompts (i) that are crossed with ratings (r'): (p:i) x r', as suggested by Wolfe and Manalo (1998), if at least two essay prompts were taken by each examinee. However, it may not possible to obtain an essay score reliability coefficient that simultaneously takes into account both ratings and prompts as potential sources of score error for a single-prompt TOEFL CBT essay test.

Future Avenues for TOEFL Essay Scoring

1. Automatic Essay Scoring: E-rater™

There have been a series of serious research efforts at ETS for the past 10 years or so directed toward the development of an automatic scoring system for essays, called "e-rater™". Since January 1999, e-rater™ is being operationally used to score GMAT (Graduate Management Admission Test) essays as a second reader. More recently, the e-rater™ system has been experimented with TWE and TOEFL CBT essays (Burstein & Chodorow, 1999; Sheehan & Burstein, 2000). More research seems to be necessary at this point in order for the e-rater to be used for the scoring of non-native speakers' essays.

but it appears clear that the new technology will dramatically change the nature of TOEFL essay scoring, if adopted in the future.

2. Integrative Writing Tasks: New TOEFL

The New TOEFL (or TOEFL 2000) is under development and scheduled to be launched sometime in the near future. The integrative writing tasks (such as reading-writing, listening-and-writing) are likely to be used, possibly with a stand-alone writing task similar to the current CBT writing or TWE tasks (Cumming, Kantor, Powers, Santos, & Taylor, 2000). The use of multiple writing tasks will surely enhance the consequential validity and reliability of TOEFL essay tests and bring about positive wash-back effects for classroom ESL/EFL (English as a Second/Foreign Language) instruction. However, it could also raise new challenges for scoring due to the built-in dependencies between the receptive skills (i.e., listening and reading) and writing. For this reason, it will become critically important to examine the potential implications of new writing tasks for the essay scoring, score reliability, and fairness of the New TOEFL.

References

Berstein, J., Chodorow, M. (June, 1999). *Automated Essay scoring for Nonnative English Speakers.* Paper presented at the Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, Maryland.

Breland, H., Muraki, E., & Lee, Y. (2000). *Comparability of TOEFL-CBT writing prompts for different prompt assignment and gender groups.* Draft Report. Princeton, NJ: ETS.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series. 18). Princeton, NJ: ETS.

DeMauro, G. (1992). *An investigation of the appropriateness of the TOEFL test as matching variable to equate TWE topics* (TOEFL research report 37). Princeton, NJ: ETS.

Educational Testing Service. (1997). *Test of Written English guide* (Fourth Ed.). Princeton, NJ: ETS.

Educational Testing Service. (1998). *Computer-based TOEFL score user's guide.* Princeton, NJ: ETS.

Educational Testing Service. (2000). *TOEFL CBT Information Bulletin.* Princeton, NJ: ETS.

Golub-Smith, M., Reese, C. & Steinhaus, K. (1993). *Topic and topic type comparability on the Test of Written English* (TOEFL Research Report 42). Princeton, NJ: ETS.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6(1),* 83-102.

Mislevy, R. J. (1992*). Linking educational assessments: Concepts, issues, methods, and prospects (Plicy Issue Perspective).* Princeton, NJ: ETS.

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the TWE* (TOEFL Research Report No. 40). Princeton, NJ: ETS.

Powers, D. & Kubota, M. (1998). *Qualifying readers for the online scoring network: Scoring argument essays* (RR-98-28). Princeton, NJ: ETS.

Sheehan, K. S., & Berstein, J. (2000). *An analysis of e-rater scoring for TOEFL CBT essays.* Draft report. Princeton, NJ: ETS.

Wolfe, E. W., & Manalo, J. (1999). *An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study.* Draft Report. Princeton, NJ: ETS.

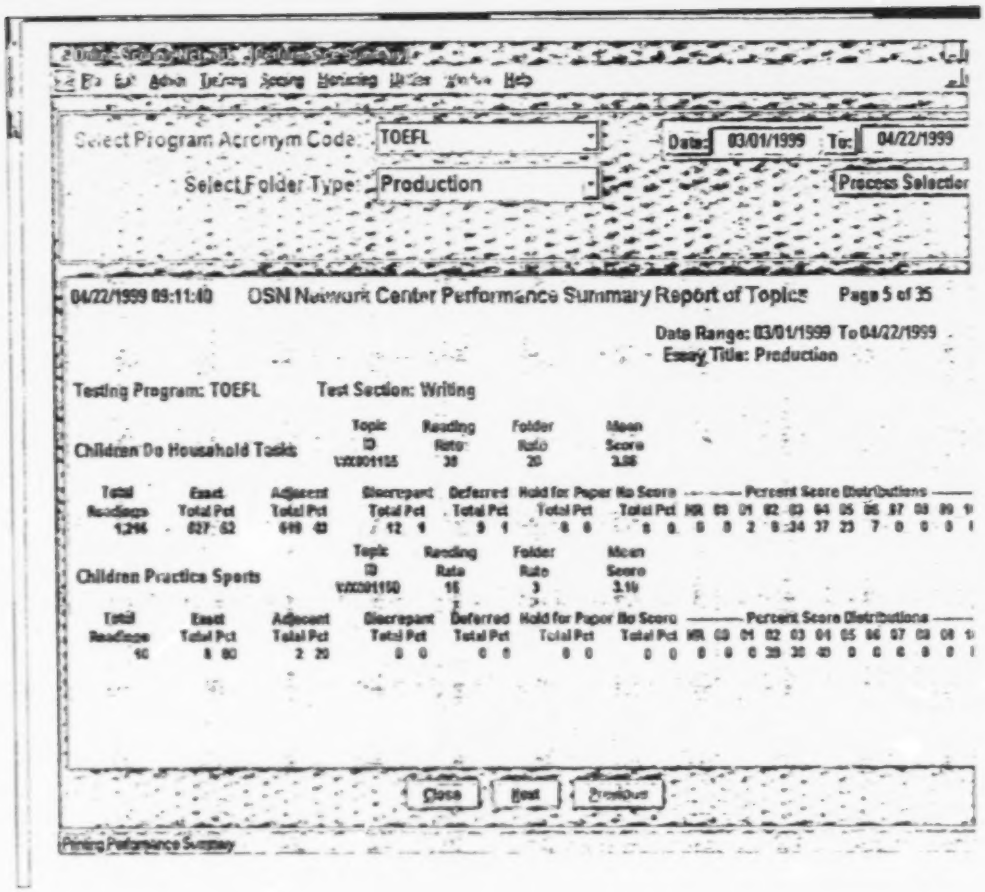
Select Program Acronym Code: TOEFL
Date: 03/01/1999 To: 04/22/1999
Select Folder Type: Production
Process Selection
04/22/1999 09:11:40 OSN Network Center Performance Summary Report of Topics Page 5 of 35
Data Range: 03/01/1999 To 04/22/1999
Essay Title: Production
Testing Program: TOEFL Test Section: Writing
Children Do Household Tasks
Children Practice Sports
Close
Print
Previous

Figure 1. OSN Performance Summary Report

BEST COPY AVAILABLE

Table 1. Inter-reader reliability for TOEFL CBT essays for 1999

(a) Cross-tabulations of the first and second ratings (January through December 1999)

		The Second Rating						
	Frequency Percent	1	2	3	4	5	6	Total
The First Rating	1	788 0.51	497 0.32	76 0.05	12 0.01	2 0.00		1375 0.89
	2	483 0.31	3621 2.35	3231 2.10	438 0.28	20 0.01	1 0.00	7794 5.06
	3	104 0.07	3413 2.21	19070 12.37	12001 7.78	1104 0.72	46 0.03	35738 23.18
	4	10 0.01	444 0.29	12145 7.88	35985 23.34	12242 7.94	1094 0.71	61920 40.17
	5	3 0.00	36 0.02	1098 0.71	12311 7.99	16605 10.77	4653 3.02	34706 22.51
	6		4 0.00	59 0.04	1194 0.77	5013 3.25	6361 4.13	12631 8.19
	Total	1388 0.90	8015 5.20	35679 23.14	61941 40.18	34986 22.69	12155 7.88	154164 100.00

Frequency Missing = 457

Note: 1. The shaded numbers indicate the frequencies of perfect agreement between the two ratings in each score category.

2. 1st Rating: Mean = 4.029; SD=1.033, 2nd Rating: Mean = 4.022; SD= 1.031

(b) Inter-reader reliabilities and agreement rates for the first and second ratings (January through December 1999)

No. of Essays	Relative Correlation	Alpha	Absolute Perfect	Adjacent	Perfect+ Adjacent	Discrepant
154,164	0.73	0.84	53.5%	42.8%	96.3%	3.7%

BEST COPY AVAILABLE

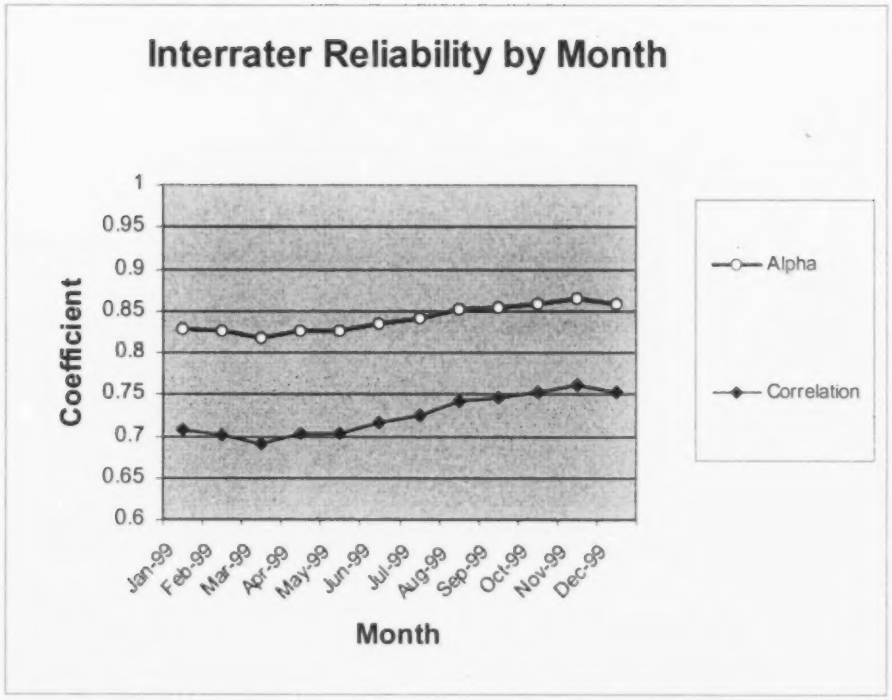

Figure 2. Interrater reliability by month for the year of 1999.

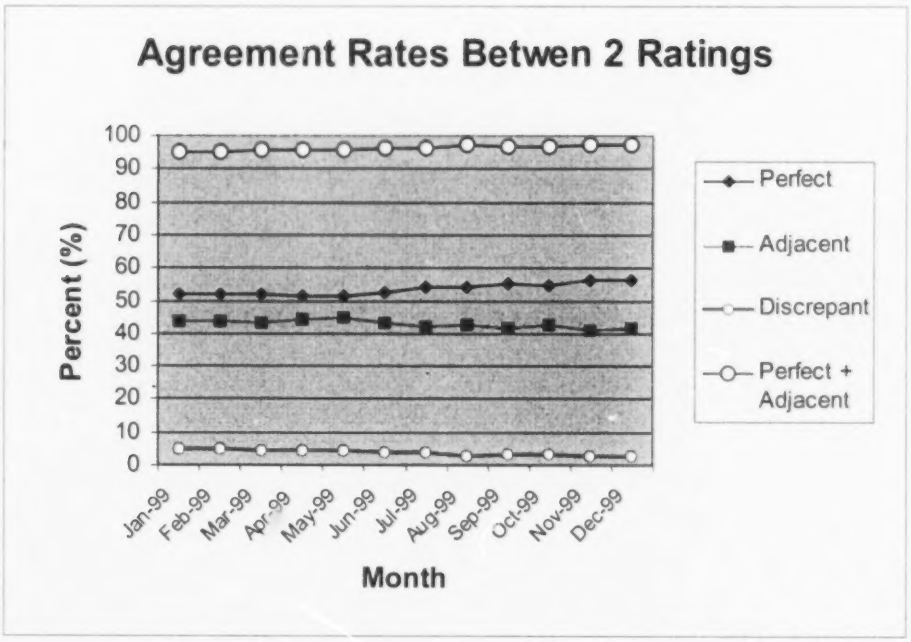

Figure 3. Agreement rates between the first and second ratings by month

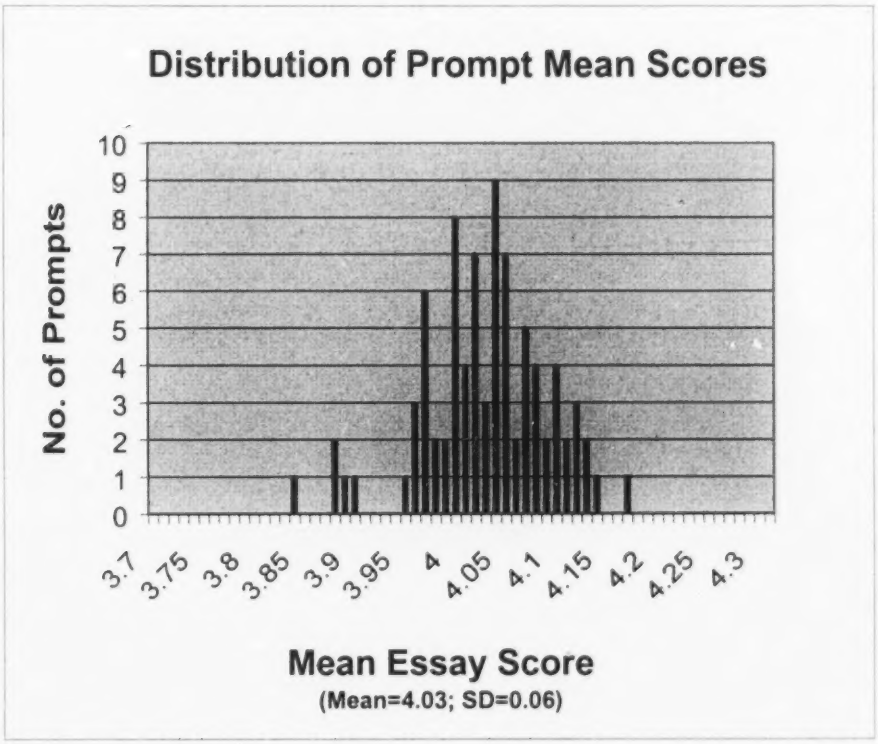

Figure 4. Distribution of prompt mean scores for 1999

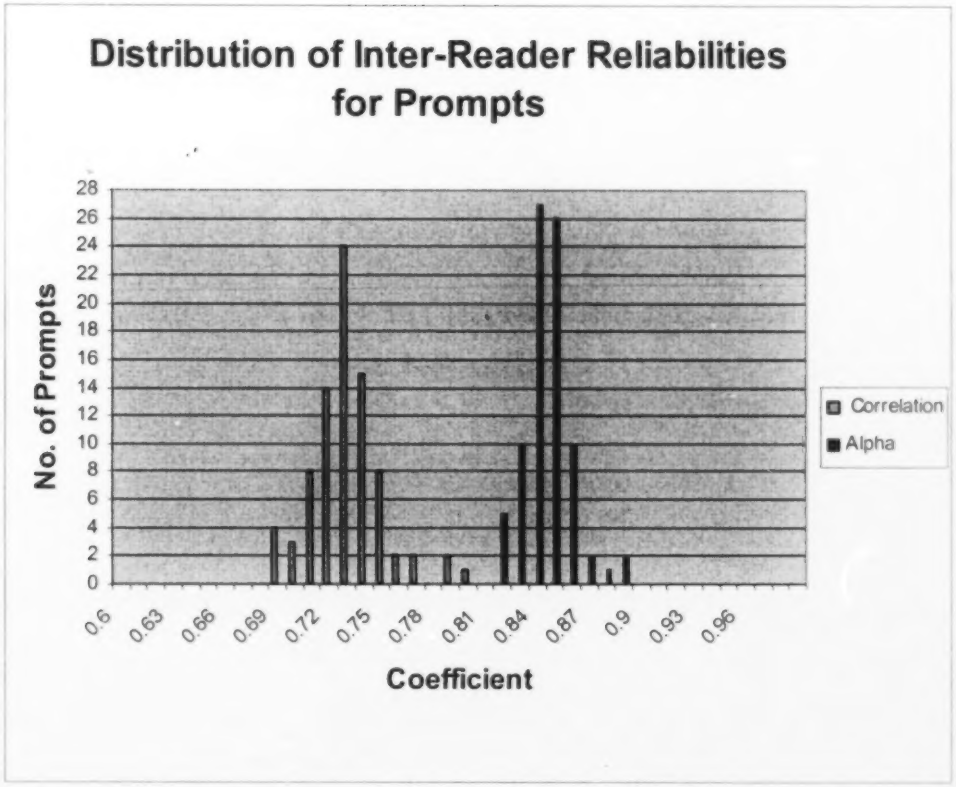

Figure 5. Distribution of inter-reader reliabilities for prompts

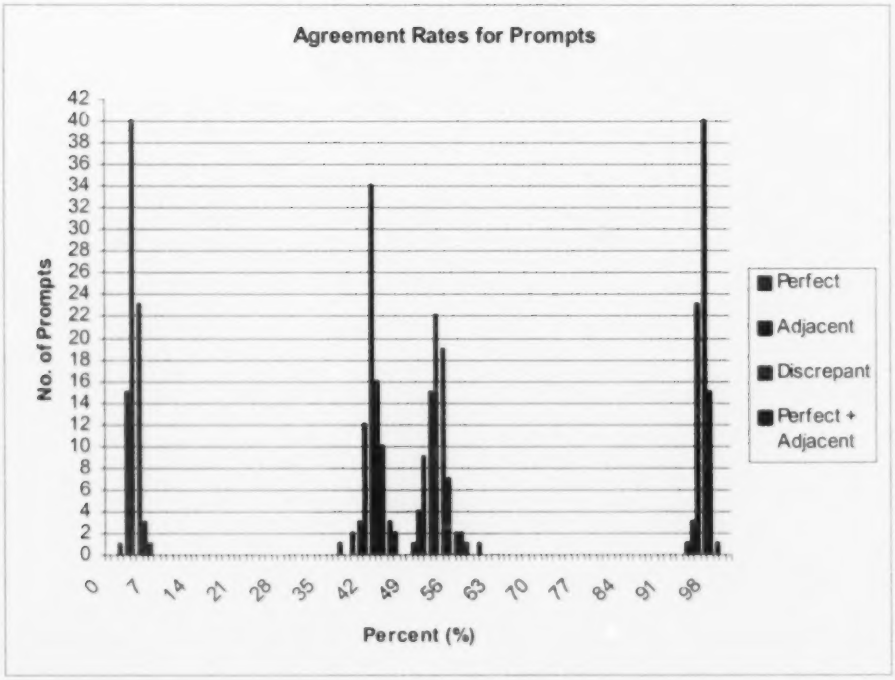

Figure 6. Distribution of inter-reader reliabilities for prompts